ПРИМЕНЕНИЕ ИНС ДЛЯ АНАЛИЗА КАЧЕСТВА БИОЧИПОВ

В. В. Горошко

ВВЕДЕНИЕ

Развитие генетики имеет огромное значение для познания природы вообще и природы человека в частности. Одна из основных задач этой области состоит в определении активных, выраженных генов организма, а также генного строения организма вообще, т.е. в секвенировании генома. Получить ответы на эти вопросы позволяет эксперимент гибридизации микроматриц – одно из последних достижений экспериментальной молекулярной биологии, позволяющее проводить параллельный анализ экспрессии десятков тысяч генов [1, 2].

Биологические микрочипы или, как их чаще называют – DNA microarrays (микроматрицы ДНК) [3].

Первостепенной задачей анализа микроматриц является определение качества слайдов. Стандартный подход, основанный на простых правилах с установкой границ допустимости для каждого параметра ячеек матрицы, не дает точного результата, поэтому эксперту приходится перепроверять каждый слайд в поисках плохих спотов, что занимает большое количество времени.

Для решения данной проблемы было предложено использовать искусственную нейронную сеть (ИНС) для обработки параметров спотов. Был разработан алгоритм формирования обучающей выборки и обученная таким образом сеть была применена для анализа качества спотов.

ПОСТАНОВКА ЗАДАЧИ

В результате анализа исходного изображения специальными программами можно получить множество параметров характеризующих каждый спот. На основании этих параметров можно оценить качество спотов, что частично реализовано в MAIA [4]. Однако стандартный подход, основанный на простых правилах с установкой границ допустимости для каждого параметра, не дает точного результата, поэтому исследователю/эксперту приходится перепроверять каждый слайд в поисках плохих спотов (расплывшиеся и слившиеся споты) и артефактов (царапины, ворсинки, «кляксы»).

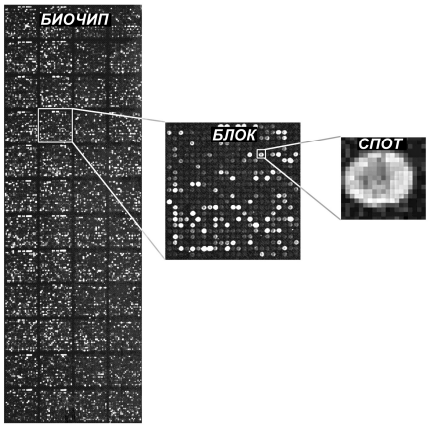

Рис. 1. Общая структура отсканированного слайда биочипа

Идея заключалась в том, чтобы обучить ИНС на множестве обработанных экспертом данных, отличать хорошие споты от плохих. При этом на вход подавать параметры спота. С выхода сети снимался параметр качества спота – действительный ($\in[0..1]$).

На вход программы, написанной в среде MATLAB, подавались файлы формата *.txt полученные при помощи программы MAIA 2.75. Такой файл представляет собой таблицу, колонки в которой разделены табуляцией. Каждый файл содержал 25392 строк и представлял собой скан биочипа состоящего из 48 блоков размером 23 на 23 ячейки. Соответствующая структура изображена на рисунке 1.

АЛГОРИТМ ФОРМИРОВАНИЯ ОБУЧАЮЩЕЙ ВЫБОРКИ

В работе с ИНС пожалуй самой сложной задачей является корректное обучение сети. С целью оптимизации обучения ИНС был предложен следующий алгоритм формирования обучающей выборки. Схема представлена на рисунке 2.

Создание выборки можно разделить на несколько этапов.

1. Выбирается несколько слайдов, по которым мы будем обучать нашу сеть. В нашем случае из 12 слайдов мы выбрали 8 для обучения и оставили остальные 4 в качестве тестовых.

2. Из слайдов выбираем только колонки с необходимыми нам параметрами и формируем одну большую таблицу.

3. Разделяем нашу таблицу на две выборки «SPOT», которая содержит параметры тех ячеек, в которых есть пятно, и соответственно «NO_SPOT» – параметры пустых ячеек.

4. Что бы избежать опасности переобучения эти 2 выборки были прорежены. То есть были отсеяны похожие пятна. В качестве меры схожести было выбрано эвклидово расстояние в пространстве параметров:

$$\sqrt{(x_1-x_2)^2+(y_1-y_2)^2+\ldots}<\delta. \qquad (1)$$

5. Формирование итоговой выборки с помощью генератора случайных чисел заполняем итоговую выборку данными из 2 массивов.

В процессе исследований были проведены исследования на двух различных типах обучающих выборок.

1. Выборка, содержащая одинаковое количество пустых ячеек и ячеек с пятнами.

2. Выборка, в которой соотношение ячеек с пятнами к пустым было близким к такому соотношению на реальном слайде (на слайдах на 25392 в среднем приходится 3843 ячеек со спотами).

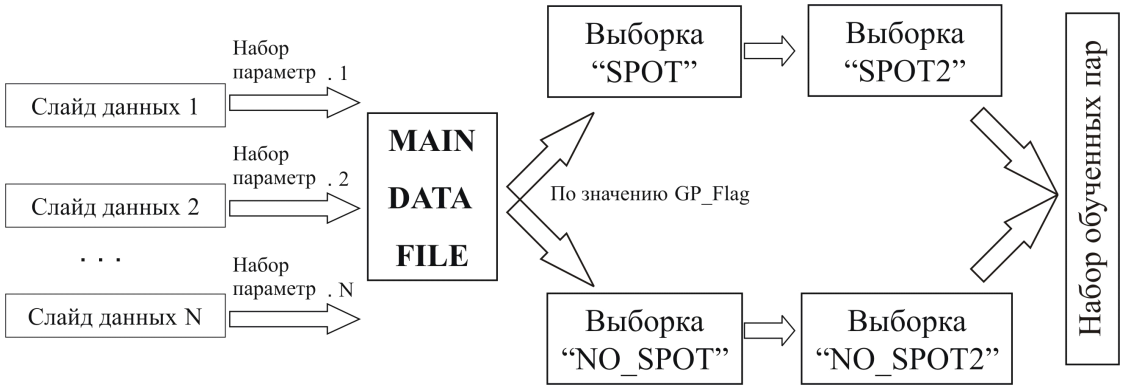

Рис. 2. Схема формирования обучающей выборки

Выборка типа 1 оказалась более эффективной.

Сформировав, таким образом, выборку из порядка трех тысяч элементов, было проведено обучение нейронной сети. Сеть – двухслойный персептрон с 15 нейронов в каждом скрытом слое. Количество нейронов в слое подбиралось вручную, путем наращивания сети в ходе эксперимента. В качестве алгоритма обучения был выбран метод обратного распространения ошибки в модификации Левенберга-Марквардта [5].

При подаче на вход слайда в 25392 строк, обученная нейронная сеть ошибается только на 30–150 ячейках (0,1–0,6% ошибок), что позволяет экономить время в течении эксперимента. Так же симуляции ИНС проходит быстрее, чем анализ качества слайдов по, тем же параметрам посредством алгоритмических методов.

Недостатком данного метода, является зависимость от сторонних программных продуктов и также необходимость некоторой последующей обработки.

Литература

1. *Brazma A., Vilo J.* Gene expression data analysis // FEBS Letter. 2000. V. 480. P. 17–24.
2. *Kurella M., Li-Li Hsiao, Takumi Yoshida et. al.* DNA microarray analysis of complex biologic processes // Journal of the American Society of Nephrology. 2001. V. 12, P. 1072–1078.
3. *Мирзабеков А. Д.* Биочипы в биологии и медицине XXI века / Вестник РАН. М., 2003.
4. *Novikov E.* MAIA – Micro Array Image Analysis. Version 2.75. User manual. Institute Curie. 2005–2006.
5. *Bishop C. M.* Neural Networks for Pattern Recognition. Oxford, Clarendon Press, 1997.